# 网络舆情推文的热度测度模型构建*

■ 黄微 刘熠 许烨婧 孙悦
吉林大学管理学院 长春 130022

摘要：[目的/意义]数据获取是网络舆情研究的第一个阶段，在大量数据面前，构建网络舆情推文热度测度模型能够快速筛选出能为网络舆情研究所用的数据。[方法/过程]借鉴信息论中平均自信息量的定义，使用层次分析法与Haker News排名算法构建网络舆情热度测度模型。[结果/结论]通过在微博抓取数据，计算得出针对该数据集的热度阈值，验证该热度测度模型的准确度。事实证明，网络舆情推文热度测度模型能够很好地完成推文热度的计算，并且能够达到较高的计算准确率。

关键词：网络舆情 推文热度 层次分析
分类号：G250 G206
DOI：10.13266/j.issn.0252-3116.2019.20.002



## 1 引言

随着互联网技术的发展，网络推文数量呈指数级增长，各种自媒体平台每日都会产生海量的推文信息。在网络舆情研究的过程中，如果将互联网中所有的舆情推文数据一次性抓取，将会带来数据的灾难。如何在数据获取阶段，有选择性有目的性地抓取可能造成舆情事件的高热度数据，屏蔽低热度数据，是网络舆情信息获取阶段的一个难题。目前网络舆情热度的研究可分为单个时间点的热度研究与时间段内的热度趋势研究两个方面。在单个时间点的热度研究中，以网络舆情推文中转发数、评论数、点赞数、粉丝数等一系列定量数据为指标的网络舆情推文评价体系构建为主，例如梁昌明等将推文附加属性分为博主特征热度影响力、内容特征热度影响力、传播特征热度影响力和受众特征热度影响力，进而构建了微博热度评价指标体系[1]。杜慧等利用因果模型从文章数量、点击量、评论量、来源数量几方面来描述主题热度，并且把时间作为关键变量考虑在内[2]。在时间段的网络舆情趋势研究当中，徐旖旎利用马尔科夫链，对不同时间点下微博定量数据的关系进行分析，绘制了舆情热度曲线图，并对舆情的趋势进行了预测[3]。黄微等通过将微博转发评论数按时间的走势分析，对微博舆情的老化度进行了计算[4]。总体来说目前网络舆情热度研究多是针对于转发数、评论数、点赞数、粉丝数等一系列定量数据的研究，而对网络舆情推文内容本身的研究不多，缺乏网络舆情推文内容与作者影响力对推文热度影响的讨论。

本文从推文内容研究出发，结合推文定量数据与作者定量数据的热度计算，综合考虑了时间对推文热度的影响因素，引用了信息论中平均自信息量的概念与热度计算中的Haker News排名算法，建立了网络舆情推文热度测度模型，根据热度高低进行筛选，实现了网络舆情抓取过程的初步过滤。该模型的提出弥补了当前网络舆情热度测度模型研究中对推文内容考虑欠缺的不足，同时对于推文的热度计算不仅考虑点赞、转发与评论数量等推文附加信息的多少，还增加了作者网龄与推文存续时长等作为时间维度，考虑单位时间内的点赞、转发与评论即附加信息的速率这一因素对推文热度的影响。

## 2 推文热度测度模型构建

### 2.1 推文热度及自信息量的概念

2.1.1 推文热度概念 推文热度表示在微博微信等自媒体平台中，作者发表的文章、图片、视频受到关注、讨论、传播的程度。在网络舆情信息获取的过程

* 本文系国家自然科学基金面上项目“大数据环境下多媒体网络舆情信息的语义识别与危机响应研究”（项目编号：71473101）研究成果之一。
作者简介：黄微（ORCID：0000-0003-0448-9563），教授，博士生导师；刘熠（ORCID：0000-0002-7360-2091），博士研究生；许烨婧（ORCID：0000-0003-1128-2878），博士研究生；孙悦（ORCID：0000-0001-7343-8343），硕士研究生。
收稿日期：2019-03-07 修回日期：2019-06-21 本文起止页码：17-25 本文责任编辑：易飞

中,通过热度计算筛选出在单位时间内更受到关注、传播范围更广泛、讨论更频繁的推文,这些推文,很有可能作为潜伏中的网络舆情,更容易引发轰动;同时过滤掉受关注度低、传播范围窄、不经常被讨论的推文,防止在网络舆情研究中出现数据灾难。本文将推文热度量化为 0 至 1 之间的一个数,以对其进行描述。

2.1.2 自信息量的概念 根据香农信息论的相关概念,事件集合 X 中事件 x = a 的自信息量定义为:

$$I_X(a_i) = -logP_X(a_i) \quad \text{公式(1)}$$

其中 $0 \leqslant P_X(a_i) \leqslant 1$ 代表事件 x = a 发生的概率,且 $\sum_{i=1}^{n} P_X(a_i) = 1$。

自信息量表示事件发生前,事件的不确定性,同时表示事件发生后,事件所包含的信息量。事件自信息量越高,事件的不确定性越高,反之,事件的确定性就越高。

本文考虑到网络舆情推文信息包含推文文字信息、视频信息与图像信息,难以在不涉及语义识别的前提下提取多媒体信息的语义,因此采用推文文字信息内容、视频标题与图像标题作为研究对象。统计网络舆情推文的关键词词频,将关键词集合理解为事件集合 X,将某个关键词理解为事件 x = a,某个关键词出现的频率理解为 $P_X(a_i)$,从而可以计算出特定词的自信息量。

本文采用计算平均自信息量的方式计算单个推文的信息量:

$$AvgI_k = \frac{1}{n} \sum_{i=0}^{n} -logP_X(a_i) \quad \text{公式(2)}$$

其中 k 为推文的编号,n 为该第 k 条推文中所含的有用词的个数,$a_i$ 为第 k 条推文中出现的第 i 个有用词,$P_X(a_i)$ 为该有用词在所有推文中出现的频率。

## 2.2 层次分析模型及判断矩阵的构建

2.2.1 层次分析模型构建 层次分析模型建立的目的在于,用网络舆情推文的各项定量数据来描述网络舆情推文的热度。利用层次分析法,构造判断矩阵,能够尽可能减少不同指标之间相互比较的困难,提高准确度,从而有效计算出各指标间的权重关系。本文建立了如图 1 所示的层次分析模型。模型的目的是计算网络舆情推文热度,即模型的目标层。根据马尔科夫·格拉德威尔提出的流行三要素理论,物体想要流行必须具备流行的基本要素,即关键人物法则、环境威力法则和内容附着力法则,在借鉴文献[1,5-6]对网络舆情热度的影响因素的基础上,本文认为,网络舆情推文热度受到作者影响力、内容感染力和网络传播力三个中间层指标共同影响,。其中作者影响力包含作者粉丝增长率、作者发文增长率、作者关注增长率三个因素层指标;内容感染力包含内容丰富度、平均自信息量两个因素层指标;网络传播力包含推文被转发加权速率、推文被评论加权速率与推文被点赞加权速率三个因素层指标。

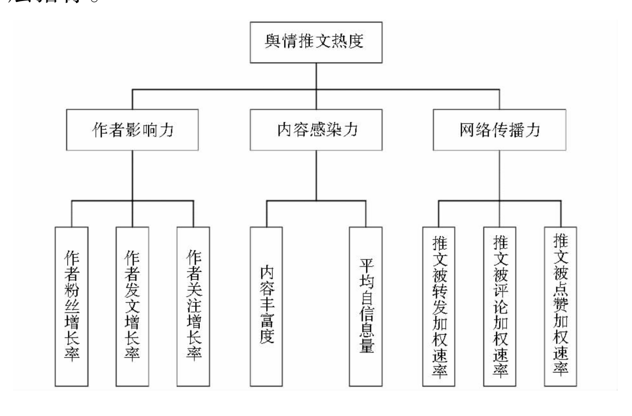



图 1 网络舆情推文热度层次分析模型

(1)作者影响力指作者在自媒体平台中的影响范围与活跃程度。作者发表的推文,在该作者人际范围内传播,人际范围越广、越活跃的作者发表的推文,传播速度越快,造成推文热度越高。包括三个底层指标:①作者的粉丝增长率,反映出作者的影响范围的辐射程度,粉丝增长率越快,影响范围覆盖速率越快,未来可能关注此推文的人数就会越多,进而造成推文的热度增加。②作者发文增长率,可以作为作者在自媒体平台活跃程度的衡量标准,一定程度上反映出作者的影响力程度,发文增长率越高,证明作者在该平台越活越,会正向影响作者所发推文的热度。③作者的关注增长率,也是作者活跃程度的衡量标准,一定程度上反映出作者的影响力,也会正向影响作者所发推文的热度值。作者粉丝增长率、发文增长率与关注增长率均来自网络爬虫抓取的粉丝数、发文数与关注数以及作者创建自媒体账户至作者发表此推文的时间跨度的计算结果,具体计算过程将在后续内容中体现。

(2)内容感染力指的是推文作者所发表推文的质量与吸引力。质量越高的推文被他人关注的程度就会越高,例如含有图片、视频的推文就会比纯文字的推文更吸引人。内容感染力具有两个底层指标:①内容丰富度,该指标根据网络爬虫抓取的推文内容部分计算,统计推文的词数与是否含有视频情况,得到内容丰富度这一定量指标,具体计算过程见后续章节;②平均自信息量,代表的是该推文或者类似推文在自媒体平台

中出现的频率,平均自信息量越高,说明该推文讨论内容出现频率越低,反之该推文讨论内容出现频率越高。如果一个类型的推文经常出现在自媒体平台当中,我们可以认为该类型的推文是目前的热议话题,这样的推文被关注的程度会相对较高,具体计算过程详见后续内容。

(3)网络传播力指的是用户发表的推文的传播速度与互动能力。用户的转发行为最能够反映出用户对于该推文的参与度与传播水平,文献[5]认为,评论、点赞与转发行为均能体现用户互动参与程度,用户互动参与度越高,继而转发扩散该推文的可能性就越大,从而影响到该推文的总体热度。本文设计将网络传播力分为推文被转发加权速率、推文被评论加权速率与推文被点赞加权速率三个底层指标,三个底层指标来源于网络爬虫抓取的转发数、评论数、点赞数与推文发表至被抓取的时间跨度的加权计算结果。关于加权,本文设计了一个理想值,作为排除推文作者的粉丝参与推文传播的权重,以弱化网络传播力与作者粉丝数的相关性。本理想值由经验得来,数值为0.75,关于该理想值的进一步界定、改进与完善,作者将另外行文进行研究。

2.2.2 判断矩阵构建　网络舆情推文热度层次分析模型建立后,需要比较各指标的相对重要性,从而建立层次模型的判断矩阵。记网络舆情推文热度为A、作者影响力为$B_1$、内容感染力为$B_2$、网络传播力为$B_3$、作者粉丝增长率为$C_1$、作者发文增长率为$C_2$、作者关注增长率为$C_3$、内容丰富度为$C_4$、推文平均自信息量为$C_5$、推文被转发加权速率为$C_6$、推文被评论加权速率为$C_7$、推文被点赞加权速率为$C_8$。

通过文献调研与专家调查法,作者影响力$B_1$比内容感染力$B_2$稍微重要,网络传播力$B_3$比作者影响力$B_1$稍微重要,网络传播力$B_3$比内容感染力$B_2$较强重要。作者粉丝增长率$C_1$比作者发文增长率$C_2$稍微重要,作者粉丝增长率$C_1$比作者关注增长率$C_3$强烈重要,作者发文增长率$C_2$比作者关注增长率$C_3$较强重要。平均自信息量$C_5$比内容长度$C_4$强烈重要。推文被转发加权速率$C_6$比推文被评论加权速率$C_7$稍微重要,推文被转发加权速率$C_6$比推文被点赞加权速率$C_8$弱于较强重要,推文被评论加权速率$C_7$比推文被点赞加权速率$C_8$强于同等重要。

根据表1及上述分析构建判断矩阵见图2。

构建判断矩阵后,需要检验矩阵满意一致性,一致性检验标准如公式(3)与公式(4)所示:

表1　判断矩阵标度及含义

| 因素i比因素j | 量化值 |
|---|---|
| 同等重要 | 1 |
| 稍微重要 | 3 |
| 较强重要 | 5 |
| 强烈重要 | 7 |
| 极端重要 | 9 |
| 判断矩阵性质 | $a_{ij}=\frac{1}{a_{ji}}$ |
| 两相邻判断的中间值 | 2,4,6,8 |

| A | $B_1$ | $B_2$ | $B_3$ |
|---|---|---|---|
| $B_1$ | 1 | 3 | 1/3 |
| $B_2$ | 1/3 | 1 | 1/5 |
| $B_3$ | 3 | 5 | 1 |

| $B_1$ | $C_1$ | $C_2$ | $C_3$ |
|---|---|---|---|
| $C_1$ | 1 | 3 | 7 |
| $C_2$ | 1/3 | 1 | 5 |
| $C_3$ | 1/7 | 1/5 | 1 |

| $B_2$ | $C_4$ | $C_5$ |
|---|---|---|
| $C_4$ | 1 | 1/7 |
| $C_5$ | 7 | 1 |

| $B_3$ | $C_6$ | $C_7$ | $C_8$ |
|---|---|---|---|
| $C_6$ | 1 | 3 | 4 |
| $C_7$ | 1/3 | 1 | 2 |
| $C_8$ | 1/4 | 1/2 | 1 |

图2　层次模型判断矩阵

$$CR=\frac{CI}{RI}<0.1 \qquad \text{公式(3)}$$

$$CI=\frac{\lambda_{max}-n}{n-1} \qquad \text{公式(4)}$$

其中$\lambda_{max}$是矩阵的最大特征根,$n$为矩阵的阶数,$RI$如表2所示:

表2　一致性标准 *RI* 取值规则

| 矩阵阶数 | 0 | 1 | 2 | 3 |
|---|---|---|---|---|
| *RI* | 0 | 0 | 0.58 | 0.90 |

通过计算,$CR_A=0.0214$,$CR_{B_1}=0.036$,$CR_{B_2}=0$,$CR_{B_3}=0.0102$,均满足一致性检验。

计算可得舆情推文热度层次分析模型的权重如下所示:

$$W_A=\begin{bmatrix}0.2583\\0.1047\\0.6370\end{bmatrix}\quad W_{B_1}=\begin{bmatrix}0.6491\\0.2790\\0.0719\end{bmatrix}\quad W_{B_2}=\begin{bmatrix}0.1250\\0.8750\end{bmatrix}$$

$$W_{B_3}=\begin{bmatrix}0.6250\\0.2385\\0.1365\end{bmatrix}$$

## 2.3 推文热度计算模型及阈值计算模型构建

2.3.1 推文热度计算模型　目前常用的推文附加信息热度计算模型有Reddit排名算法[7]、PageRank网页排名算法[8]、Hacker News排名算法等。其中Reddit排名算法主要针对具有支持数与反对数统计的推文附加信息热度排名计算,PageRank是针对有网页链接指向的网页热度排名计算,这两种算法并不能支持本文所收集的微博推文热度计算模型。Haker News模型参数

涉及到了发文的时间与点赞数,与本文所提及的热度计算模型契合度较高,因此本文将在 Haker News 热度计算模型的基础上进行修改,从而构建全新的推文热度计算模型。

本文基于 Haker News 排名算法,构建推文附加信息热度计算模型,表达式如公式(5)所示:

$$\mathrm{r} = \frac{(P-1)}{(t+2)^{G}} \qquad 公式(5)$$

其中 P 为推文的得票数,t 为以天为单位的时间。因此通过公式计算,越短时间内得票数越多的推文排名会越靠前,得票数一定,随时间增加,推文的排名会慢慢降低。在公式(5)中,G 为重力因子,之所以成为重力因子,是因为随着 G 的增大,推文的排名会被时间下拉得越快,通常来说设置 G = 1.8。公式中分子之所以减 1,目的是排除掉作者对推文的投票。

本文拟在 Haker News 的基础上对热度排名算法进行修改,利用其时间影响排名的思路重新构建热度计算模型。根据上文层次分析模型,作者影响力对应的指标有作者粉丝数(Funs#)、作者发文数(Publications#)、作者关注数(Follows#),在此还要考虑作者创建微博至今的时间跨度($t_{person}$)。网络传播力对应的指标有推文转发数(Forward#)、推文评论数(Comment#)和推文点赞数(ThumbUp#),在此我们也考虑到推文发布至今的时间跨度($t_{article}$)。内容感染力对应的指标有是否包含视频(*Video*)、推文有效字数(*Word*#)和推文平均自信息量(SelfInformation#)、理想值(ρ = 0.75)。

本文分别应用 Haker News 排名算法结合上节计算的权重,计算出作者影响力($\mathrm{r}_{person}$)与网络传播力($\mathrm{r}_{article}$)。计算公式如下:

$$\mathrm{r}_{person} = \frac{(0.649\ 1 \times \mathrm{Funs\#} +\ 0.279\ 0 \times \mathrm{Publications\#} + 0.071\ 9 \times \mathrm{Follow\#})}{(\frac{t_{person}}{24 \times 60} + 2)} \qquad 公式(6)$$

$$\mathrm{r}_{article} = \rho \times \frac{(0.625\ 0 \times \mathrm{Forward\#} + 0.238\ 5 \times \mathrm{Comment\#} + 0.136\ 5 \times \mathrm{ThumbUp\#}\ )}{(\frac{t_{article}}{24 \times 60} + 2)^{1.8}} \qquad 公式(7)$$

通过将一部视频看作 100 有用词,将推文平均自信息量与字数按照公式(8)进行标准化,对内容充实度热度的计算公式如公式(9)所示:

$$\mathrm{g(x)} = \frac{x - x_{min}}{x_{max} - x_{min}} \qquad 公式(8)$$

$$\mathrm{r}_{content} = \begin{cases} 0.875\ 0 \times (1 - g(\mathrm{SelfInformation}^{\#})\ )\ + 0.1250 \times g(Word^{\#}),\ Video = False \\ 0.875\ 0 \times (1 - g(\mathrm{SelfInformation}^{\#})) + 0.125\ 0 \times g(Word^{\#} + 100),\ Video = True \end{cases} \qquad 公式(9)$$

在计算作者影响力、网络传播力与内容感染力的基础上,首先进行数据标准化处理如公式(10)所示,将三种热度结果限制在 0 到 1 之间,考虑到推文影响力比网络传播力与内容感染力表述推文热度高低的能力更强,因此将推文影响力加权,计算总体热度公式如公式(11)所示:

$$f(x) = \frac{x}{1+x} \qquad 公式(10)$$

$$\mathrm{r} = 0.258\ 3 \times f(\mathrm{r}_{person} + 0.104\ 7 \times f(\mathrm{r}_{content}) + 0.637\ 0 \times f(\mathrm{r}_{article}) \qquad 公式(11)$$

2.3.2 推文热度阈值计算模型 为了更精确地计算推文热度的阈值,本文建立了如下数学模型:

$$f(r_k) = \begin{cases} 1, \&r_k \geqslant x \\ 0,\ else \end{cases} \qquad 公式(12)$$

$$Err = \frac{\sum(\mathrm{Target}_k - \mathrm{f}(\mathrm{r}_k))^2}{\mathrm{N}} \qquad 公式(13)$$

其中 $\mathrm{r}_k$ 代表第 *k* 条推文的热度,x 推文热度的阈值,*N* 代表推文总数。优化该数学模型,选取合适的 x 值,使得 *Err* 最小。

2.3.3 关键词库与敏感词库 本文设置两种类型的词库:关键词库与敏感词库[9],以达到算法查漏补缺的作用。在现实处理中,热度达标的推文中可能存在一些对网络舆情研究相关度不大的推文,例如天气预报、心灵鸡汤、抽奖活动等。这些推文往往都携带辨识度很高的关键词,在热度达标推文中,过滤掉带有这些关键词的推文,能够提高热度判断的准确度。同理,在热度不达标的推文中,可能存在一些与舆情演进相关度很高的推文,例如还未引起轰动的地方性事件、含有敏感信息的推文等。这些推文在网络舆情研究中,不能忽视,因此可以对热度不达标的推文采用敏感词筛选的方式,提取出含有特定敏感词的推文,从而提高热度判断的准确度。

在本文研究中,敏感词来源于 *CSDN* 用户整理的开放下载的“2017 敏感词库”,包含了有关色情、暴力、反动、贪腐、民生等类别的敏感词,用于从热度不达标的推文中筛选需要抓取的推文。关键词库包括作者针

对微博推文信息整理的有关天气、抽奖、广告等类别的关键词，用于在热度达标的推文中过滤掉与网络舆情研究不相关的推文。

### 2.4　推文热度测度流程及评价模型

2.4.1　推文热度测度流程　在层次模型的基础上，结合热度计算模型，本文应用图3所示推文热度测度流程，对网络舆情信息的热度进行测度。热度计算模型镶嵌在层次模型中，用于对舆情推文热度进行整体全面的热度计算。在完成推文热度计算的基础上，利用推文热度阈值计算模型，估算出热度阈值标准，将在阈值以上的推文视作热度达标推文，在阈值以下的推文视作热度不达标推文。通过关键词过滤与敏感词筛选，最终确定热度达标的网络舆情推文与热度不达标的网络舆情推文。

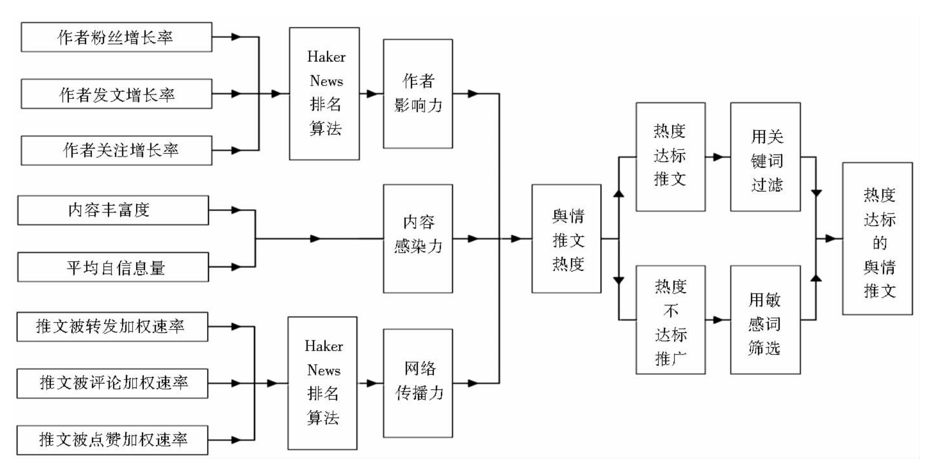



图3　网络舆情热度测度模型

2.4.2　推文热度测度评价模型　本文应用准确率与召回率对推文热度测度模型进行评价，设实际热度达标的推文集合为T，不达标的为～T，计算热度达标的推文集合为CT，计算热度不达标的推文集合为～CT，则准确率为ACC，召回率为REC，计算公式为：

$$ACC = \frac{|T \cap CT| + |\sim T \cap \sim CT|}{N} \qquad \text{公式(14)}$$

$$REC = \frac{|T \cap CT|}{T} \qquad \text{公式(15)}$$

准确率与召回率反映了两个不同方面的指标，准确率反映了整体计算的准确性，召回率表示热度计算达标的集合占实际热度达标集合的比率，衡量是否漏报。

## 3　数据获取与数据清理

### 3.1　数据获取

本文的数据来源于新浪微博，通过八爪鱼网络爬虫工具，抓取500名博主的70 833条微博数据作为初始数据。数据包含了作者微博名、推文获取时间、推文发布时间、推文内容、推文转发数、推文评论数、推文点赞数、是否包含视频、作者创建微博时间、作者关注数、作者粉丝数、作者发文数12列信息，部分数据如图4所示。通过数据过滤与数据清理，过滤掉重复数据，清理掉非法字符总计获得63 816条微博数据。

| 作者 | 作者关注数 | 作者粉丝数 | 作者发文数 | 作者创建微博时间 | 内容 | 转发 | 评论 | 点赞 | 发布时间 | 收集时间 | 是否含有视频 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 如来福lf696696 | 2345 | 306 | 442 | 2012/8/10 10:22 | 今天我的「微博等级」又升一级，记录一下我在微博的成就O微博 | 0 | 0 | 0 | 2017/4/25 17:38 | 2019/4/10 11:29 | FALSE |
| 如来福lf696696 | 2345 | 306 | 442 | 2012/8/10 10:22 | 能拿奖品这种事一定要和大家说一说的，否则以后还怎么在一起愉 | 0 | 0 | 0 | 2017/4/25 17:39 | 2019/4/10 11:29 | FALSE |
| 天添钱官方微博 | 72 | 374 | 939 | 2014/11/13 15:03 | 发布了头条文章：《网贷风险备付金存废定论：可以设置风险缓释 | 0 | 0 | 0 | 2017/4/25 17:57 | 2019/4/10 22:04 | FALSE |
| 有缘就相逢 | 77 | 90 | 457 | 2016/8/24 19:41 | 中国诗词欣赏第四首! | 0 | 0 | 0 | 2017/4/25 19:31 | 2019/4/9 14:33 | FALSE |
| 灵异老朱 | 79 | 160246 | 621 | 2017/4/20 20:28 | 世界之大无奇不有，你相信诡异的存在吗？背后呼的冷风一吹，你 | 2 | 3 | 11 | 2017/4/25 20:08 | 2019/4/10 16:27 | FALSE |
| 离地两米的一楼 | 125 | 45 | 637 | 2016/11/29 23:47 | 入戏太深了，现在的小盆友真的好过分哦 哈哈哈L秒拍视频 #搞 | 0 | 0 | 0 | 2017/4/25 23:10 | 2019/4/10 17:06 | TRUE |
| 有缘就相逢 | 77 | 90 | 457 | 2016/8/24 19:41 | 第5首，从军行是古乐府旧题，以此为名，李白、王维、杨炯、王 | 0 | 0 | 0 | 2017/4/26 7:38 | 2019/4/9 14:33 | FALSE |
| 啰嗦地板 | 0 | 0 | 0 | 2011/2/23 2:04 | @本真的世界 | 0 | 0 | 0 | 2017/4/26 7:44 | 2019/4/10 20:21 | FALSE |
| kjjvvdnpyjsyx | 148 | 62 | 740 | 2015/10/29 14:32 | 你想找的东西就在你知道的地方，我不会再找不到东西了眼镜一个 | 0 | 0 | 1 | 2017/4/26 9:51 | 2019/4/10 11:20 | FALSE |
| 文刀公随 | 0 | 0 | 0 | 2015/12/25 14:23 | //@现货大鳄-老于:可喜可贺！！！ | 0 | 0 | 0 | 2017/4/26 11:18 | 2019/4/10 15:25 | FALSE |
| 山下清泉 | 56 | 49 | 648 | 2010/12/5 14:36 | 发表了一篇转载博文《[转载]小学里的"成人式腐败"有多严重？》 | 0 | 0 | 0 | 2017/4/26 11:50 | 2019/4/10 16:44 | FALSE |
| 山下清泉 | 56 | 49 | 648 | 2010/12/5 14:36 | 计生干部调查失独者12年 被骂猪狗不如会遭报应 @辛辣头条 （分 | 0 | 0 | 0 | 2017/4/26 12:02 | 2019/4/10 16:44 | FALSE |
| lucky哆蕾咪 | 255 | 44 | 203 | 2017/4/24 22:09 | 全国被曝光很多例了，每次看都触目惊心，希望引起大家关注，社 | 0 | 0 | 0 | 2017/4/26 13:43 | 2019/4/9 10:36 | FALSE |
| lucky哆蕾咪 | 255 | 44 | 203 | 2017/4/24 22:09 | 在土耳其伊斯坦布尔的街头，有一种逆天的自动贩卖机，好心人把 | 0 | 0 | 0 | 2017/4/26 13:46 | 2019/4/9 10:36 | FALSE |
| 纪问题儿童易飘扬 | 3825 | 1047 | 369 | 2011/7/12 6:37 | 4？ | 0 | 0 | 0 | 2017/4/26 13:56 | 2019/4/9 14:10 | FALSE |
| 天添钱官方微博 | 72 | 374 | 939 | 2014/11/13 15:03 | 发布了头条文章：《天添钱2017年劳动节放假公告》 °天添钱2017 | 0 | 0 | 0 | 2017/4/26 18:00 | 2019/4/10 22:04 | FALSE |
| 离地两米的一楼 | 125 | 45 | 637 | 2016/11/29 23:47 | 一些关键的税改细节包括：放弃支持边境调节税、对美国企业的离 | 1074 | 1691 | 1774 | 2017/4/26 21:53 | 2019/4/10 17:06 | FALSE |
| 离地两米的一楼 | 125 | 45 | 637 | 2016/11/29 23:47 | 海军要继续努力呀? | 0 | 0 | 0 | 2017/4/26 22:04 | 2019/4/10 17:06 | FALSE |
| 离地两米的一楼 | 125 | 45 | 637 | 2016/11/29 23:47 | 漂亮呀 | 0 | 0 | 0 | 2017/4/26 22:28 | 2019/4/10 17:06 | FALSE |
| 离地两米的一楼 | 125 | 45 | 637 | 2016/11/29 23:47 | 怀旧时光，周星驰错过朱茵终生不娶，最后的谢谢让人心酸 ♡ L秒 | 0 | 0 | 0 | 2017/4/26 22:47 | 2019/4/10 17:06 | TRUE |
| 离地两米的一楼 | 125 | 45 | 637 | 2016/11/29 23:47 | 这算是永动机吗 | 0 | 0 | 0 | 2017/4/26 23:06 | 2019/4/10 17:06 | FALSE |
| TIENS/小亮 | 208 | 245 | 414 | 2013/3/19 9:34 | 这老外活该挨揍 | 0 | 0 | 0 | 2017/4/26 23:19 | 2019/4/9 19:20 | FALSE |
| 有缘就相逢 | 77 | 90 | 457 | 2016/8/24 19:41 | 第6—7首，阅读其实就是一个积累过程，机会都是给有心人准备的 | 0 | 0 | 0 | 2017/4/27 7:38 | 2019/4/9 14:33 | FALSE |
| 肥皂泡也有梦想 | 207 | 285 | 748 | 2010/6/26 0:00 | 生活想要给我更多之前，一定会考验我一番，生而为人，不容易， | 0 | 0 | 1 | 2017/4/27 9:14 | 2019/4/10 20:39 | FALSE |
| 离地两米的一楼 | 125 | 45 | 637 | 2016/11/29 23:47 | 【垃圾克星来了！这种虫子能吃塑料】西班牙和英国科研人员发现， | 622 | 1120 | 2403 | 2017/4/27 9:18 | 2019/4/10 17:06 | FALSE |

图4　初始数据截图示例

由于应用新的数据集数据量过大,不能够完全运用人工进行标注,本文标注采用 TensorFlow 中开源的自然语言处理工具与 GitHub 中成型的开源卷积神经网络分类工具,先对数据进行正负向情感判断,将负面信息标记为 1,将正面情感与中立情感标记为 0,而后将负面信息中,点赞、转发、评论量过少的推文标记为 0,最后人工对标记为 1 的少量数据再次进行筛选,剔除涉及到天气类型、心灵鸡汤类型、民生福利类型、政府公告类型、花边新闻类型、动物宠物类型、作秀类型、抽奖类型等与网络舆情研究内容不符的推文,记为不抓取推文。最终获得不抓取推文 51 531 条,标记为 0,需要抓取的推文共计 12 285 条,标记为 1。

### 3.2 中文分词

本文采用 Java 编程技术,从 Maven 中导入中文分词工具 ansj,针对推文内容,编程实现了中文分词。考虑到例如代词、介词、语气词等一系列与推文本身语义无关的特殊词语,本文采用 ansj 中的停用词工具,定向抓取。在本次实验中只采用分词结果中的一般名词、人名、音译人名、地名、音译地名、机构团体名、其他专名、名词性惯用语、名词性语素、新词、处所词、一般动词、副动词、动名词、动词性语素、形容词、副形词、名形词、副词、区别词这 20 类词。并添加“有”“没有”“还”“是”“也”等未滤出的停用词。

在分词与停用词设置的基础上,对已分词数据进行词频统计,计算单个词出现的频率,由于该频率数值过小,在 Java 中采用 BigDecimal 包函数对频率进行精准表述与计算。具体分词过程与自信息量计算代码如下:

```
//词语词数的 HashMap
HashMap < String, Integer > map = new HashMap < String, Integer >
( );
//分词结果数组
ArrayList < Result > ls = new ArrayList < Result > ( );
for ( int i = 0; i < sheet. getRows( ); i + + ) {
  //分词核心代码
  String cellinfo = sheet. getCell(3, i). getContents( );
  Result strs = ToAnalysis. parse( cellinfo);
  ls. add( strs);
  for( java. util. Iterator < Term > itr = strs. iterator( ) ; itr. hasNext
  ( ); ) {
    Term temp = itr. next( );
    //停用词过滤
    if( expectedNature. contains( temp. getNatureStr( ) )
    && ! stopWords. contains( temp. getNatureStr( ) ) ) {
      String tempString = temp. getName( );
      if( map. containsKey( tempString) ) {
      map. put( tempString, map. get( tempString) + 1);
      total + +;
      } else{
      map. put( tempString, 1);
      total + +;
      }
    }
  }
}
for ( Entry < String, Integer > entry: list) {
  String key = entry. getKey( );
  Integer value = entry. getValue( );
  //利用 BigDecimal 精准计算词频
  BigDecimal bvalue = new BigDecimal( value);
  BigDecimal btotal = new BigDecimal( total);
  BigDecimal percentage = bvalue. divide ( btotal, 10, RoundingMode.
  HALF_UP);
  //计算自信息量
  double selfInformation = - Math. log( percentage. doubleValue( ) )/
  Math. log(2);
}
```

## 4 实验与结果

### 4.1 实验过程

4.1.1 分词与平均自信息量计算 在 Java 环境下利用 ansj 中文分词工具,可以实现如图 5 所示的中文分词结果。在分词的基础上,利用 HashMap 记录每个词与其出现的次数,并根据公式(1)计算该词的自信息量,如图 6 所示。最后统计每个词的自信息量,与每一条推文中的词数,根据公式(2)计算每篇推文的平均自信息量,如图 5 所示:

#湖南舆情#史上最长湘绣《千鹤图卷》被毁,
万科物业任烧一晚不报警

strs:,湖南,舆情,史,最长,湘绣,鹤,图,毁,万科,物业,
任,烧,不,报警

sumSelfInfor:185.76389688434324
字数:14
avgInformation:13.268849777453088

图 5 中文分词结果及推文平均自信息量

4.1.2 热度计算 在计算推文平均自信息量后,根据推文发布时间与推文抓取时间计算单条推文发布时长,根据作者创建微博时间与推文抓取时间,计算创建微博至今的时长,时间长度均以分钟为单位,最终形成如图 7 所示的推文基本数据示例。利用公式(6) - 公式(11),可以计算得出如图 8 所示的结果。

黄微，刘熠，许烨婧，等.网络舆情推文的热度测度模型构建[J].图书情报工作,2019,63(20):17－25.



key:复活;value:46;percentage:0.0000320368;selfInfomation:14.929910423278491
key:备用;value:46;percentage:0.0000320368;selfInfomation:14.929910423278491
key:新气象;value:46;percentage:0.0000320368;selfInfomation:14.929910423278491
key:孤单;value:46;percentage:0.0000320368;selfInfomation:14.929910423278491
key:气候;value:46;percentage:0.0000320368;selfInfomation:14.929910423278491
key:母女;value:46;percentage:0.0000320368;selfInfomation:14.929910423278491
key:类型;value:46;percentage:0.0000320368;selfInfomation:14.929910423278491
key:孝子;value:46;percentage:0.0000320368;selfInfomation:14.929910423278491
key:谎言;value:46;percentage:0.0000320368;selfInfomation:14.929910423278491
key:超标;value:46;percentage:0.0000320368;selfInfomation:14.929910423278491
key:黄河;value:46;percentage:0.0000320368;selfInfomation:14.929910423278491
key:过分;value:46;percentage:0.0000320368;selfInfomation:14.929910423278491
key:丢失;value:46;percentage:0.0000320368;selfInfomation:14.929910423278491
key:开会;value:46;percentage:0.0000320368;selfInfomation:14.929910423278491
key:轻生;value:46;percentage:0.0000320368;selfInfomation:14.929910423278491
key:透明;value:46;percentage:0.0000320368;selfInfomation:14.929910423278491
key:安全感;value:46;percentage:0.0000320368;selfInfomation:14.929910423278491
key:影像;value:46;percentage:0.0000320368;selfInfomation:14.929910423278491

图6　词频统计与平均自信息量

| 作者 | 作者关注数 | 作者粉丝数 | 作者发文数 | 创建时间差 | 内容 | 转发 | 评论 | 点赞 | 发布微博至今时长 | 平均自信息量 | 是否含有视频 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 认真的赵先森 | 254 | 667905 | 307 | 3783541.2 | 昨天庭审结束的 | 316 | 613 | 2293 | 472120 | 12.040494 | FALSE |

图7　推文基本数据截图示例

| 作者影响力 | 网络传播力 | 内容感染力 | 热度 |
|---|---|---|---|
| 0.9939729 | 0.9352139 | 0.3671142 | 0.8909113 |

图8　热度计算结果截图示例

### 4.2　实验结果

根据推文热度模型，对推文中定量数据进行热度计算，得到63 817条数据的热度，经过归一化处理的热度取值范围范围从0到1，目前计算热度最低推文的热度为0.000 02，最高热度推文的热度为0.935。本文截取了部分热度计算的结果予以展示，如图9高热度推文信息和图10低热度推文信息。从图9可以看到，高热度推文发布者具有较高的粉丝数、关注数与发文数，推文转发、评论、点赞数较多，平均自信息量较低，属于网络热议内容。除了少量的广告类型、心灵鸡汤类型的推文，大部分推文均被标记为需要抓取的推文。从图10中观察到，低热度推文的转发、评论、点赞数很少，大部分为0，发布者的粉丝数、关注数、发文数均不高，平均自信息量较高，属于网络中偏冷门内容。

| 作者 | 作者关注数 | 作者粉丝数 | 作者发文数 | 创建微博时间差 | 内容 | 转发 | 评论 | 点赞 | 发布微博时间差 | 是否含有视频 | 平均自信息量 | 字数 | 作者影响力 | 网络传播力 | 内容感染力 | 热度 | 是否抓取 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 蜜蜂舆情 | 371 | 18691 | 1361 | 816793 | 【21岁奥迪司机开车玩手机 | 185124 | 107397 | 167909 | 276886 | TRUE | 12.23091 | 45 | 0.956575 | 0.999893 | 0.37393 | 0.923166 | 1 |
| 文案人于极 | 2367 | 42158 | 849 | 1801850 | 听说台湾人太客气了？——| | 484241 | 361922 | 468813 | 324593 | TRUE | 11.93783 | 17 | 0.956821 | 0.999949 | 0.373353 | 0.923204 | 1 |
| 文案人于极 | 2367 | 42158 | 849 | 1801850 | 文案不读书会死吗？会死。 | 236324 | 257783 | 606655 | 242417 | FALSE | 10.89207 | 12 | 0.956821 | 0.999953 | 0.373548 | 0.923227 | 1 |
| 蜜蜂舆情 | 371 | 18691 | 1361 | 816791 | @应采儿 今年過年，我給你 | 2639 | 1733 | 12197 | 134447 | FALSE | 10.98626 | 31 | 0.956575 | 0.998694 | 0.382547 | 0.923304 | 0 |
| 易辩-李雪芹律 | 377 | 50965 | 896 | 843972 | 【求雪得雪！#京津冀迎来大 | 31 | 72 | 1324 | 79843 | TRUE | 12.71078 | 56 | 0.982676 | 0.991073 | 0.365373 | 0.923393 | 0 |
| 认真的赵先森 | 254 | 667905 | 307 | 3783540 | 发布了头条文章：《一张「 | 412 | 308 | 936 | 146616 | FALSE | 11.74037 | 20 | 0.993973 | 0.987773 | 0.357776 | 0.923414 | 1 |
| 认真的赵先森 | 254 | 667905 | 307 | 3783542 | 发布了头条文章：《写在老 | 3667 | 4716 | 26666 | 701727 | FALSE | 11.28393 | 9 | 0.993973 | 0.98706 | 0.362228 | 0.923426 | 0 |
| 易辩-李雪芹律 | 377 | 50965 | 896 | 843975 | #吃了隔夜冰西瓜 小肠坏死# | 10291 | 221575 | 832585 | 363525 | FALSE | 13.44323 | 21 | 0.982676 | 0.999835 | 0.3129 | 0.923481 | 1 |
| 认真的赵先森 | 254 | 667905 | 307 | 3783542 | #唐山黄淑芬# 经与警方核实 | 8406 | 7774 | 39329 | 681142 | FALSE | 13.02732 | 24 | 0.993973 | 0.99302 | 0.326681 | 0.9235 | 1 |
| 蜜蜂舆情 | 371 | 18691 | 1361 | 816795 | 这时候不打波广告，对不起 | 175151 | 184213 | 84499 | 373566 | FALSE | 10.60033 | 8 | 0.956575 | 0.999819 | 0.378068 | 0.923551 | 1 |
| 灵异老朱 | 79 | 160246 | 621 | 1036556 | 黄巢----杀人八百万的传说 | 926077 | 1199976 | 2250813 | 858519 | FALSE | 14.21726 | 18 | 0.99312 | 0.999887 | 0.287702 | 0.923573 | 1 |
| 灵异老朱 | 79 | 160246 | 621 | 1036556 | 喜迎十九大最为一个违背科 | 493235 | 1074529 | 1751951 | 776191 | FALSE | 13.98257 | 9 | 0.99312 | 0.999862 | 0.287937 | 0.923582 | 1 |
| 蜜蜂舆情 | 371 | 18691 | 1361 | 816792 | #基因编辑婴儿#世界首例免 | 131751 | 19841 | 314835 | 193689 | FALSE | 11.95001 | 62 | 0.956575 | 0.999929 | 0.378258 | 0.923641 | 1 |
| 认真的赵先森 | 254 | 667905 | 307 | 3783544 | 忍无可忍！#唐山黄淑芬# | 8881 | 5542 | 17376 | 723745 | FALSE | 11.53946 | 2 | 0.993973 | 0.989533 | 0.351421 | 0.92387 | 1 |
| 认真的赵先森 | 254 | 667905 | 307 | 3783544 | 下午肇事司机给我打来电话 | 2433 | 5223 | 26150 | 722358 | FALSE | 10.72447 | 20 | 0.993973 | 0.984853 | 0.382209 | 0.924112 | 1 |
| 灵异老朱 | 79 | 160246 | 621 | 1036559 | 【某涯最诡异的九个帖】据 | 17810 | 15067 | 57378 | 989544 | FALSE | 12.64806 | 27 | 0.99312 | 0.99246 | 0.338941 | 0.924207 | 0 |
| 认真的赵先森 | 254 | 667905 | 307 | 3783540 | 因法院依据执行笔录裁定刘 | 395 | 533 | 1402 | 129092 | FALSE | 13.27626 | 53 | 0.993973 | 0.992039 | 0.339627 | 0.924231 | 1 |
| 认真的赵先森 | 254 | 667905 | 307 | 3783541 | 黄淑芬说：「我买房，我还 | 364 | 540 | 1119 | 179759 | TRUE | 11.43825 | 29 | 0.993973 | 0.984256 | 0.388133 | 0.924352 | 1 |
| 易辩-李雪芹律 | 377 | 50965 | 896 | 843975 | #审判实务#【在生产经营地 | 444416 | 424637 | 216365 | 351355 | FALSE | 13.67793 | 42 | 0.982676 | 0.999934 | 0.321188 | 0.924412 | 1 |
| 蜜蜂舆情 | 371 | 18691 | 1361 | 816790 | #晚安# 真正能自律的人，可 | 126236 | 128547 | 21755 | 107873 | FALSE | 10.61917 | 22 | 0.956575 | 0.999971 | 0.38578 | 0.924456 | 1 |
| 认真的赵先森 | 254 | 667905 | 307 | 3783540 | 今天执行局向我下达了国家 | 87 | 250 | 1461 | 108714 | FALSE | 12.45335 | 49 | 0.993973 | 0.98941 | 0.358185 | 0.9245 | 0 |
| 灵异老朱 | 79 | 160246 | 621 | 1036550 | 恐怖片《妈妈》 | 1512810 | 1137532 | 1918380 | 440179 | FALSE | 13.51618 | 2 | 0.99312 | 0.999973 | 0.296767 | 0.924577 | 1 |
| 蜜蜂舆情 | 371 | 18691 | 1361 | 816790 | #百度舆情#今日，小编在某 | 93209 | 195765 | 243785 | 115467 | FALSE | 10.95132 | 38 | 0.956575 | 0.999973 | 0.387386 | 0.924625 | 1 |
| 灵异老朱 | 79 | 160246 | 621 | 1036549 | 确认身份 | 0 | 15 | 0 | 76 | FALSE | 13.2098 | 2 | 0.99312 | 0.998642 | 0.305821 | 0.924677 | 0 |
| 文案人于极 | 2367 | 42158 | 849 | 1801849 | 华为荣耀手机，找《新喜剧 | 379707 | 470515 | 230244 | 59635 | TRUE | 11.32034 | 17 | 0.956821 | 0.999997 | 0.387725 | 0.92474 | 1 |
| 认真的赵先森 | 254 | 667905 | 307 | 3783540 | 车损赔偿案因黄上诉，定于 | 5747094 | 6189580 | 856 | 106154 | FALSE | 13.72369 | 10 | 0.993973 | 0.999999 | 0.296564 | 0.924793 | 1 |
| 灵异老朱 | 79 | 160246 | 621 | 1036557 | 杀人烹尸只因肚子饿？震惊 | 1521227 | 1928719 | 1027139 | 974782 | FALSE | 13.78145 | 16 | 0.99312 | 0.999892 | 0.299369 | 0.924798 | 1 |
| 认真的赵先森 | 254 | 667905 | 307 | 3783541 | 黄淑芬要起诉这些曾为我发 | 1106 | 695 | 1583 | 208007 | FALSE | 12.46681 | 46 | 0.993973 | 0.990266 | 0.355943 | 0.92481 | 1 |

图9　高热度推文信息截图示例

在热度计算的基础上，对公式(12)与公式(13)求极值，得出当阈值取0.4时，函数取得最小值为0.059，见图11。

当选取阈值为0.4时，根据公式(14)与公式(15)计算得出热度计算准确率为94%，召回率为91%。这说明在低热度群体中，还有大量的推文存在抓取的必要性。因此在低热度集合中，利用敏感词筛选，筛选出131条推文数据。在高热度集合中利用关键词过滤，过滤掉1 128条推文数据。通过筛选与过滤操作，热度计算的准确率上升至95%，召回率上升至92%。

本文算法在准确率方面略逊于卷积神经网络等机器学习算法，但是在时间复杂度方面具有明显的优势。

| 作者 | 作者关注数 | 作者粉丝数 | 作者发文数 | 创建微博时间差 | 内容 | 转发 | 评论 | 点赞 | 发布微博时间差 | 是否含有视频 | 平均自信息量 | 字数 | 作者影响力 | 网络传播力 | 内容感染力 | 热度 | 是否抓取 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| lmmingming | 0 | 0 | 0 | 3536767 | @李妮静 | 0 | 2 | 0 | 545735 | FALSE | 16.86851 | 1 | 0 | 0.008025 | 0.178477 | 0.023799 | 0 |
| 啰嗦地板 | 0 | 0 | 0 | 4275022 | 你哪种类型？@Fa軻 @第三 | 0 | 0 | 0 | 3306426 | FALSE | 15.60673 | 2 | 0 | 0 | 0.227959 | 0.023867 | 0 |
| 啰嗦地板 | 0 | 0 | 0 | 4275020 | 牛b//@学徒王自健:下巴不贝 | 0 | 0 | 0 | 1798878 | FALSE | 15.65949 | 5 | 0 | 0 | 0.228805 | 0.023956 | 0 |
| 不高兴的臼齿 | 0 | 0 | 0 | 754354 | 我在这里 2宁波·宁波诺丁汉· | 0 | 0 | 0 | 750301 | FALSE | 15.65633 | 5 | 0 | 0 | 0.228913 | 0.023967 | 0 |
| 不高兴的臼齿 | 0 | 0 | 0 | 754354 | 我在这里 2宁波·武陵大厦 | 0 | 0 | 0 | 732270 | FALSE | 15.58929 | 3 | 0 | 0 | 0.229514 | 0.02403 | 0 |
| lmmingming | 0 | 0 | 0 | 3536769 | 荒野，独行。 | 0 | 0 | 1 | 2482268 | FALSE | 15.58106 | 2 | 0 | 1.53E-04 | 0.228897 | 0.024063 | 0 |
| 不高兴的臼齿 | 0 | 0 | 0 | 754351 | 我在这里 2宁波·本雅明咖啡 | 0 | 0 | 0 | 633946 | FALSE | 15.5787 | 3 | 0 | 0 | 0.229874 | 0.024068 | 0 |
| 啰嗦地板 | 0 | 0 | 0 | 4275018 | 空气净化器 | 0 | 0 | 0 | 1154812 | FALSE | 15.5276 | 2 | 0 | 0 | 0.230802 | 0.024165 | 0 |
| 不高兴的臼齿 | 0 | 0 | 0 | 754352 | 什么牌子的卫生巾最好？ | 0 | 0 | 0 | 674830 | FALSE | 15.51555 | 3 | 0 | 0 | 0.232132 | 0.024304 | 0 |
| lmmingming | 0 | 0 | 0 | 3536769 | 这思路 | 0 | 0 | 0 | 2713564 | FALSE | 15.45347 | 1 | 0 | 0 | 0.232586 | 0.024352 | 0 |
| 文刀公随 | 0 | 0 | 0 | 1730943 | //@现货大鳄-老于:可喜可贺 | 0 | 0 | 0 | 1028408 | FALSE | 15.57692 | 6 | 0 | 0 | 0.232648 | 0.024358 | 0 |
| ABEYY | 0 | 0 | 0 | 4406222 | 冷血的中国人。 | 0 | 0 | 0 | 3929338 | FALSE | 15.47129 | 2 | 0 | 0 | 0.232841 | 0.024378 | 0 |
| 啰嗦地板 | 0 | 0 | 0 | 4275018 | 这词，难怪老爹爱看，一套 | 0 | 0 | 0 | 1147149 | FALSE | 15.51498 | 4 | 0 | 0 | 0.233066 | 0.024402 | 0 |
| ABEYY | 0 | 0 | 0 | 4406223 | 王者富民，霸者富士，仅存 | 0 | 0 | 0 | 4237543 | FALSE | 15.70067 | 12 | 0 | 0 | 0.233607 | 0.024459 | 0 |
| 天平小生 | 0 | 0 | 0 | 3118763 | 脚步印证辉煌 | 0 | 0 | 0 | 1297218 | FALSE | 15.47282 | 3 | 0 | 0 | 0.233666 | 0.024465 | 0 |
| 文刀公随 | 0 | 0 | 0 | 1730939 | 万箭齐发 航母请留步吧 | 0 | 0 | 0 | 528750 | FALSE | 15.51538 | 5 | 0 | 0 | 0.233961 | 0.024496 | 0 |
| 不高兴的臼齿 | 0 | 0 | 0 | 754353 | 我在这里 2宁波·镇海鼓楼广 | 0 | 0 | 0 | 704087 | FALSE | 15.45284 | 4 | 0 | 0 | 0.235269 | 0.024633 | 0 |
| 天平小生 | 0 | 0 | 0 | 3118762 | 好样的，冰花交警蜀黍！ | 0 | 0 | 0 | 629110 | FALSE | 15.38638 | 3 | 0 | 0 | 0.23675 | 0.024788 | 0 |
| 啰嗦地板 | 0 | 0 | 0 | 4275013 | 改名了改名了 | 0 | 0 | 0 | 261445 | FALSE | 15.32419 | 2 | 0 | 0 | 0.238049 | 0.024924 | 0 |
| 埃埃怪 | 84 | 117 | 227 | 3570949 | //@来去之间://@杜长军://@ | 0 | 0 | 0 | 753027 | FALSE | 18.66094 | 4 | 0.055314 | 0 | 0.101697 | 0.024935 | 0 |
| 啰嗦地板 | 0 | 0 | 0 | 4275022 | 快三~~~呼呼力亚！！~ | 0 | 0 | 0 | 3517891 | FALSE | 15.31711 | 2 | 0 | 0 | 0.238295 | 0.02495 | 0 |
| 天平小生 | 0 | 0 | 0 | 3118763 | 棒棒哒 | 0 | 0 | 0 | 1110357 | FALSE | 15.28355 | 1 | 0 | 0 | 0.238571 | 0.024978 | 0 |
| 不高兴的臼齿 | 0 | 0 | 0 | 754351 | 胖子的假期 2宁波·慈吉小学 | 0 | 0 | 0 | 650509 | FALSE | 15.33743 | 4 | 0 | 0 | 0.239359 | 0.025061 | 0 |
| 啰嗦地板 | 0 | 0 | 0 | 4275024 | 亲爱的，拜拜啦，呵~~~ | 0 | 0 | 0 | 3589986 | FALSE | 15.2725 | 2 | 0 | 0 | 0.239839 | 0.025111 | 0 |
| 不高兴的臼齿 | 0 | 0 | 0 | 754354 | 今年春夏上海时装周上的古 | 0 | 0 | 0 | 739831 | FALSE | 15.3563 | 6 | 0 | 0 | 0.240455 | 0.025176 | 0 |
| 曾勇0818 | 66 | 129 | 566 | 3281511 | 七休哥 | 0 | 0 | 0 | 1920532 | FALSE | 20.45337 | 1 | 0.097496 | 0 | 2.26E-05 | 0.025186 | 0 |
| 文刀公随 | 0 | 0 | 0 | 1730944 | 春天的诗或一些灵感的碎片 | 0 | 0 | 0 | 1550920 | FALSE | 15.23979 | 4 | 0 | 0 | 0.242746 | 0.025415 | 0 |
| lmmingming | 0 | 0 | 0 | 3536771 | 新年吉星照，如意祥云绕！ | 0 | 0 | 0 | 2763796 | FALSE | 15.2861 | 6 | 0 | 0 | 0.242895 | 0.025431 | 0 |

图 10　低热度推文信息截图示例

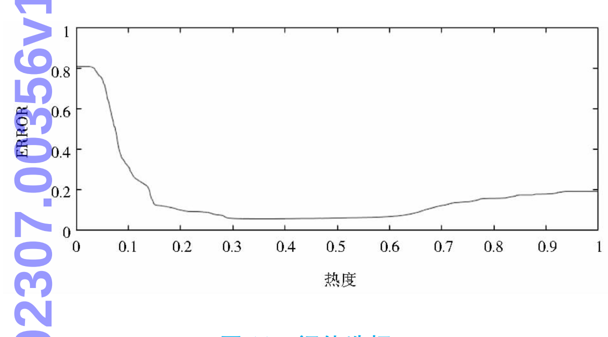



图 11　阈值选择

在数据预处理的基础上，本算法的时间复杂度为 O(N)，属于线性阶，然而涉及到自然语言处理等机器学习算法的时间复杂度均以 O(N^2)为起点。因此，本文作为网络舆情研究数据获取阶段的初步过滤算法，具有明显的速度优势。本文实验中，在 63 816 条数据中，共过滤掉 50 865 条数据，收集到 12 951 条有网络舆情研究价值的数据，在准确率 95% 的基础上总共过滤了 79% 的无用数据，性能良好。

## 5　总结与展望

本文通过层次分析模型与热度计算模型构建了完善的热度测度模型，通过对推文作者、推文内容与推文附加信息的数据计算与分析，得出推文的热度，在热度计算的基础上，通过关键词筛选与敏感词过滤，提高了热度测度的准确性。通过热度的计算，为后续网络舆情并发获取中信息的过滤提供了技术与数据支持。本文是静态的热度测度计算模型，可以在不同时间节点重复本文算法，以对动态变化的推文热度进行研究。

本文仅对微博中的推文进行了抓取与分析，有待对数据进行补充，同时推文传播加权速率中权重设置的相关研究也有待完善。在本文的基础上，今后可以扩展到多媒体网络舆情信息中图片、视频、音频信息的热度测度计算。

**参考文献：**

[1] 梁昌明，李冬强. 基于新浪热门平台的微博热度评价指标体系实证研究[J]. 情报学报，2015，34(12)：1278－1283.

[2] 杜慧，郭岩，范意兴，等. 基于因果模型的主题热度计算与预测方法[J]. 中文信息学报，2016，30(2)：50－55.

[3] 徐旖旎. 基于微博的媒体奇观网络舆情热度趋势分析[J]. 情报科学，2017，35(2)：92－97，125.

[4] 黄微，王洁晶，赵江元. 微博舆情信息老化测度研究[J]. 情报资料工作，2017(6)：6－11.

[5] 何跃，蔡博驰. 基于因子分析法的微博热度评价模型[J]. 统计与决策，2016(18)：52－54.

[6] 饶浩，文海宁. 采用实时线性模型的微博话题预警分析[J]. 图书情报工作，2017，61(15)：130－137.

[7] GLENSKI M，PENNYCUFF C，WENINGER T. Consumers and curators：browsing and voting patterns on reddit[J]. IEEE transactions on computational social systems，2017，4(4)：196－206.

[8] BERKHIN P. A survey on pagerank computing[J]. Internet mathematics，2005，2(1)：73－120.

[9] 邓一贵，伍玉英. 基于文本内容的敏感词决策树信息过滤算法[J]. 计算机工程，2014，40(9)：300－304.

**作者贡献说明：**

黄微：论文框架设计与整体思路把关；
刘熠：模型构建与论文撰写；
许烨婧：数据采集与实验；
孙悦：数据采集与论文校对。

黄微，刘熠，许烨婧，等.网络舆情推文的热度测度模型构建[J].图书情报工作,2019,63(20):17-25.

## The Construction of Heat Assessment Model for Tweets of Network Public Opinion

Huang Wei　Liu Yi　Xu Yejing　Sun Yue
School of Management, Jilin University, Changchun 130022

**Abstract:** [**Purpose/significance**] Data Collection is the first step of the study of Network Public Opinion. The construction of Heat Assessment Model for Tweets of Network Public Opinion will rapidly screen useful data over dramatic number of data. [**Method/process**] This paper cites the definition of Average Self-Information, applies Analytic Hierarchy Process (AHP) and Haker News Ranking Algorithm to construct a Heat Assessment Model for Tweets of Network Public Opinion. [**Result/conclusion**] Through the calculation of data collected from Weibo, this paper obtains the threshold of this data set. Then this paper tests the accuracy of the model, which proves this model could achieve the heat calculation precisely.

**Keywords:** network public opinion　heat of Tweets　AHP